IDS 705 Final Report

Shopee – Price Match Guarantee

Match products with descriptions and images

Team Members:

Suzy Anil
Isha Singh
Alisa Tian
Dingkun Yang

Team Number: 04

Abstract

A competitive feature amongst retail platforms is product matching which allows companies to offer products at rates competitive to other retailers selling similar products. There are many methods that combine deep learning and traditional machine learning methods to analyze image and text information to calculate the similarity between products, however, there is little research comparing the effectiveness of integrating multimodal data (product images and descriptions) under this domain (Łukasik et al., 2021). Here, we compare the performance of both unimodal and multimodal models. We trained separate models for text (SBERT and DistilBERT) and images (ResNet50 and MobileNet); the DistilBERT and ResNet50 models outperform the other two in terms of F1 score and accuracy. The multimodal model used joint embeddings from DistilBERT and MobileNet to predict product labels, which outperformed both unimodal implementations.

Introduction

Within the last decade, shopping online has become a common practice for customers worldwide. According to Oberlo, a competitor of Shopee, "the global e-commerce growth rate for 2023 is forecast at 10.4%, bringing global e-commerce sales worldwide to $6.3 trillion."[1] With an online platform, it is much easier to compare prices, search for deals and ask for a price match when ordering online. However, this search is difficult to do manually considering product

[1] https://www.oberlo.com/statistics/global-ecommerce-sales-growth

details (titles, descriptions, and images) have no standardized format within and across platforms. The same product can be listed under multiple postings with different titles and descriptions making it difficult to compare under the same platform. Retailers like Shopee and Amazon, sell products from various sellers who comply with price parity agreements that require them to adjust their prices to stand at or below other marketplaces (Kaspien, 2022). However, without a proper product matching model in place, sellers could be under-selling their products or customers could be driven away due to the greater prices. The more accurate the model is, the more competitive the company can be in the "price war"; thus, it enables the company to attract more users and enlarge the platform base, which eventually raises the market share, increases revenue, and becomes more competitive in the market. In this project, we will experiment with the capabilities of unimodal versus multimodal models to identify what is the optimal amount of information required to do semantic matching. Through preliminary research, we expect difficulty in capturing similarities in descriptions of different languages but we will explore the tradeoffs of various model architectures and which are most effective in generating physical and/or semantic similarities.

Background

The problem we are trying to tackle involves entity matching, where a single product can have different image representations and text descriptions across various platforms. This makes it challenging for customers to find the best deals when product descriptions might vary for the

same product. To address this challenge, we use multiple models to process different data types, generate embeddings, and classify product matches by calculating similarities in embeddings.

A recent research paper focused on binary classification with multimodal data from Mexican e-commerce sites (Estrada-Valenciano et al., 2022). The researchers used Convolutional Neural Network (CNN) to obtain an embedding for an image, coupled with BERT (Bidirectional Encoder Representations from Transformers) to return embeddings for the text, and then fused the two to produce one joint embedding to compare products across two different platforms. Another research paper used a new model: FashionBERT to perform cross-modal text and image matching (Gao et al., 2022). This model combines the pre-trained BERT language model with products' visual representation, with a joint embedding space that compares text and image information; we have adopted this methodology to generate embeddings to pass through our multimodal model.

While significant work has been done in entity matching for multimodal data and optimization of unimodal models for product matching, there has been insufficient research in the comparison of the two types of models. Our research will leverage transfer learning with pre-trained models to find what integration of data types results in the best similarity measures.

Data

Shopee is the leading e-commerce platform in Southeast Asia and Taiwan; its platform contains products from vendors all over the world, predominantly in Singapore and Indonesia. In 2021, the company launched a Kaggle competition aimed at improving product matching algorithms to

optimize their customers' online shopping experience (Dane et al., 2021). This data set contains both images and tabular data; specifically, 32,412 images with their respective unique IDs and 34,250 rows of observations in total. As shown in Figure 1, we divided the data set into a training set and a test set in an 80:20 ratio; then split the former into a training and validation set with the same ratio.

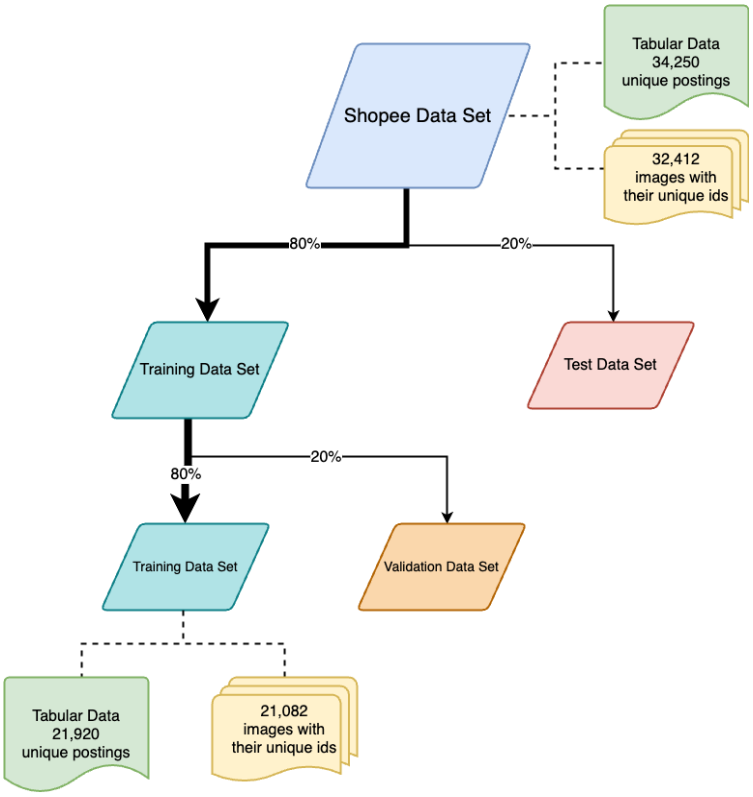


Figure 1. Flowchart for Data Splitting

With this data set, we aim to identify the products that have been posted multiple times taking into consideration that the postings may have different titles or pictures, as shown in Figure 2 below. Note that, for some posting titles, we have multiple languages other than simply English, which we would discuss in the Experiment section on how to handle this.

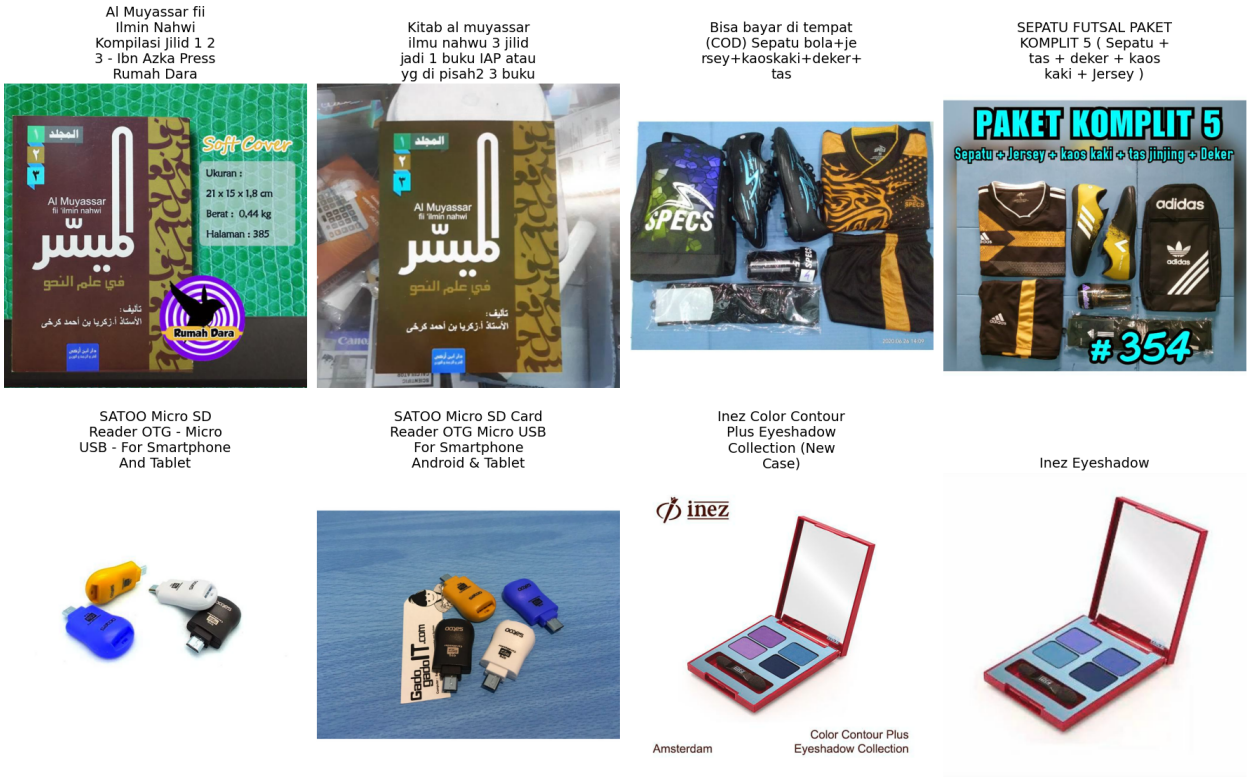


Figure 2. Four Pairs of Examples of the same products with different titles

Based on the difference between the count of unique images (21,082) and rows of observations (21,920), we would find 838 postings that use the same exact image as another posting in the training data. Another vital feature is posting titles, which is the product description for the posting. The title is case sensitive, and we have 21,920 unique posting titles, meaning 2.2% of the posting can be directly matched with each other. The mean word count in the title is around 10 words for each title. As for the target group label, there are 10,019 unique group labels, i.e. classes. As can be seen in Figure 3, more than 80% of the groups have less than 3 postings and images in them, which would cause trouble for the model to extract and learn the characteristics of each group.

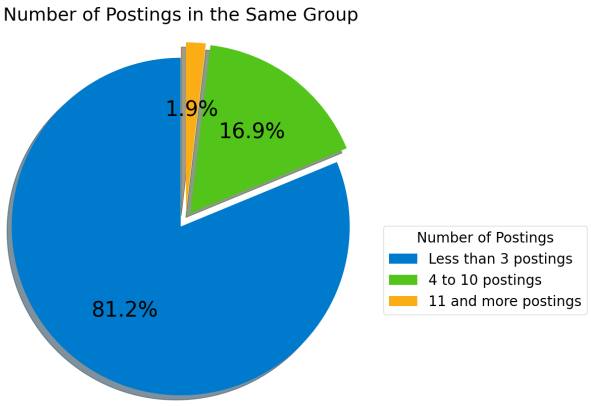


Figure 3. Pie Charts of the Number of Postings in the Same Group Respectively

To sum up, we consider the posting title and the image itself as three notable features of the data relevant to our goal of matching the same product with both text data and images. Additionally, we would treat the group label as the target.

Experiments

In our analysis, we aimed to investigate how unimodal and multimodal models would fare in the context of product matching. Intuitively, incorporating embeddings from both modalities could enhance the accuracy of product matches by providing a richer source of information. However, this approach may also introduce additional noise during the classification process, which potentially leads to overfitting the model. To understand the tradeoff, we utilize three types of analysis: text similarity, image similarity, and multimodal similarity as seen in Figure 4. Image similarity analysis involves comparing the embeddings extracted from visual features of different product images to identify matches. Text similarity analysis requires comparing the semantic and

syntactic features of product titles to generate embeddings. Multimodal similarity will use joint embeddings of a fixed length which allows for semantic alignment of product descriptions and images.

The embeddings will be passed through one final classification layer based on K-Nearest Neighbors (KNN). For each model separately, we use hyperparameter tuning with the validation dataset to see what number of neighbors will specify the best matches, and use that value for predictions on the test data (See Figures 4 & 5 in Appendix). We have 10,019 unique classes in our training data set, and within each class, we had as few as 2 images per class, which makes it difficult to find explicit similarities in product groups. With the embeddings, we expect similar products to be clustered closer together in a vector space which is how KNN predicts which class a sample belongs to.

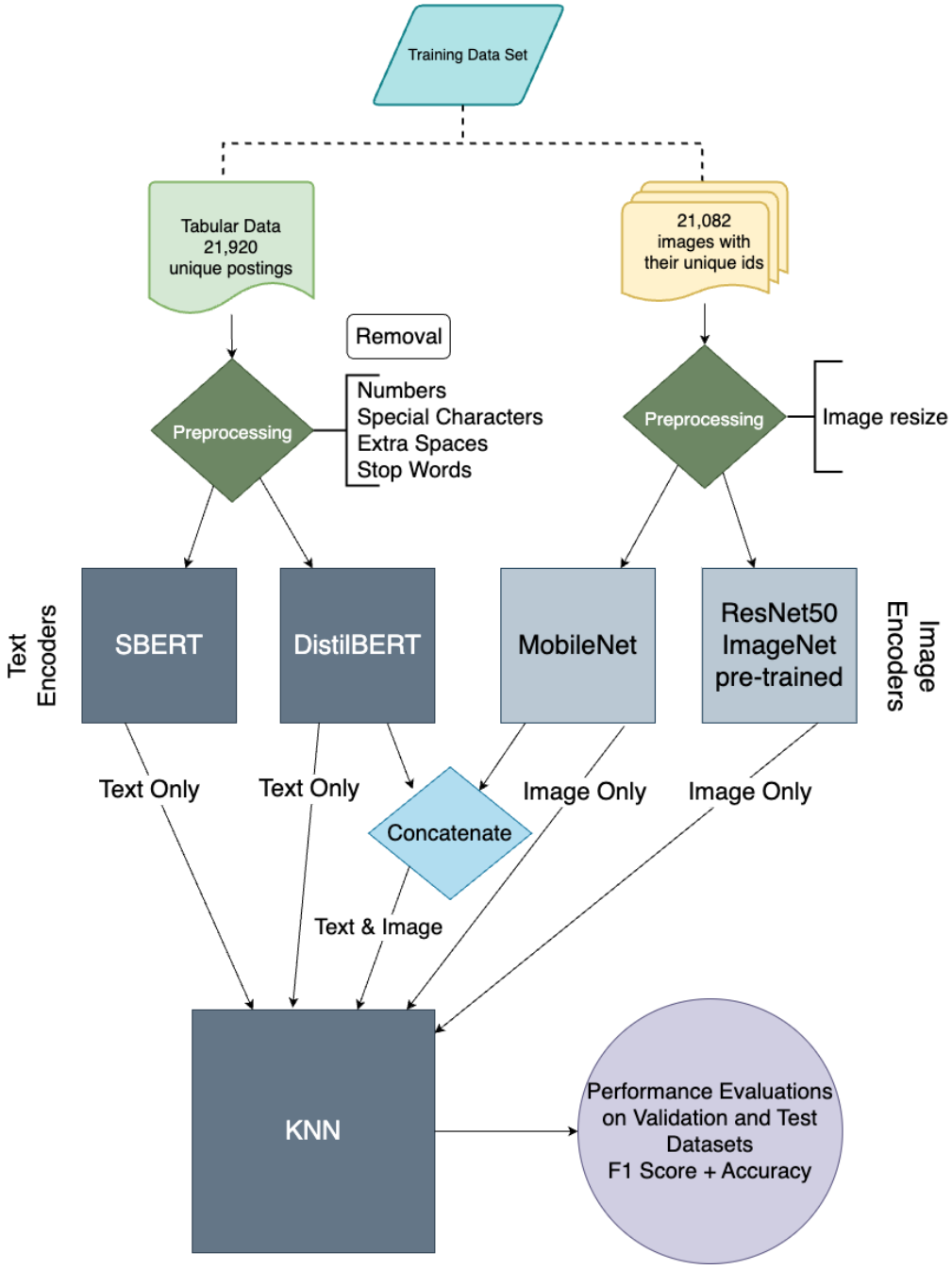


Figure 4. Methodology Flow Chart

In our preliminary analysis, we preprocessed the data to remove any inconsistencies and prepare it for encoding. For text data, we removed all numbers, special characters, and extra

spaces from the product title. Then we use the NLTK package (Bird et al., 2009) to tokenize the words, remove all stop words and convert all titles to lowercase for consistency. For image data, we first separated all the training images based on their respective group labels, and then resized them all to 224x224 pixel images to prepare them for embedding.

To obtain text and image embeddings, we will be using pre-trained models to encode our preprocessed data. The text models include a Bidirectional Encoder Representation from Transformer (BERT) as their base but include different approaches to calculating sentence similarity. The first model is a Siamese Bert (SBERT) which has two Siamese BERT encoders to compare two sentences at a time (Reimers et al., 2019). SBERT uses cosine-similarity to calculate the distance between all titles; which is appreciated, given the likelihood of many classes containing very few samples to train on. The second model is DistilBERT, which contains a distilled BERT base that utilizes fewer parameters while preserving performance (Sanh et. al., 2020). This model has been trained in various languages so it has better exposure to the diverse languages represented in our dataset. For the image models, the transfer learning models we compared shared a Deep Convolutional Neural Network (Deep CNN) architecture but applied different weights to produce embeddings. The first image model was a ResNET50 pre-trained with ImageNet weight, which has been trained on millions of images from the ImageNet database. According to Sharma et. al, (2018), ResNet50 is one of the most popular convolution neural networks for object detection and object category classification from images. For comparison, we have used MobileNet architecture as an alternative image model, inspired by Wang, et al. (2020), since it was recommended in the paper as a lightweight deep neural network with fewer parameters and higher classification accuracy.

The final multimodal model will classify based on the joint embeddings from the image and text models. Research regarding joint embedding learning has indicated multimodal similarity analysis as a powerful technique that focuses on checking the similarity between images and text (Xie et. al., 2021). The text embeddings are produced by DistilBERT and the image embeddings are from a pre-trained MobileNet model. The embeddings are concatenated, normalized, then passed through a KNN classification head. By combining both image and text embeddings, our model is able to find similarities between both modes of data and translate it to similar products on the platform.

After predictions have been made, we use F1 score and accuracy as test metrics to evaluate the image and text similarity models. In order to assess the model's general ability to find its respective class, we use accuracy as a metric. However, we wanted to make sure that the spread of classes was taken into account when discussing the model's ability to find similar samples. When calculating an F1 score, each class is rated separately so class size is considered when measuring success. Due to the sparse classes. F1 score would incorporate the imbalance of classes when measuring the success of a product match.

Results

In order to generate our own test metrics, we split the given dataset into training, validation, and test datasets by the exact process outlined in the Data section. We were able to maintain the

spread of classes so the sparsity of classes is still represented in the split datasets. This ensures consistency across all models and the embeddings the final multimodal model is fitted with.

Text Model:

Regarding our proposed text models, Table 1 illustrates that DistilBERT outperformed the SBERT model by a wide margin in all metrics. As discussed in the Experiment section, DistilBERT's superior performance can be attributed to its training in multiple languages and robust architecture.

Model	F1 Score on Validation	F1 Score on Test	Accuracy on Test
Random Guessing	0.0024	0.0023	0.0023
SBERT	0.4246	0.4291	0.4511
DistilBERT	0.4510	0.4766	0.4523

Table 1. Metrics for Text Model

Figure 5 visualizes the challenge of distinguishing clusters from text embeddings after dimensionality reduction with t-SNE. Because the groups are so specific, there could be products with embeddings that overlap but do not belong in the same label group. The largest label group in testing data contains 13 samples, which is only 0.2% of the dataset; given the minute class sizes, both text models performed significantly well. As seen in Figure 4 in the Appendix, while k increases, both models would have decreased performance on validation, which is expected. This is attributed to the sparsity in classes with more than 80% of label groups in the training data set having less than three postings which is why performance decreases if k is greater than three.

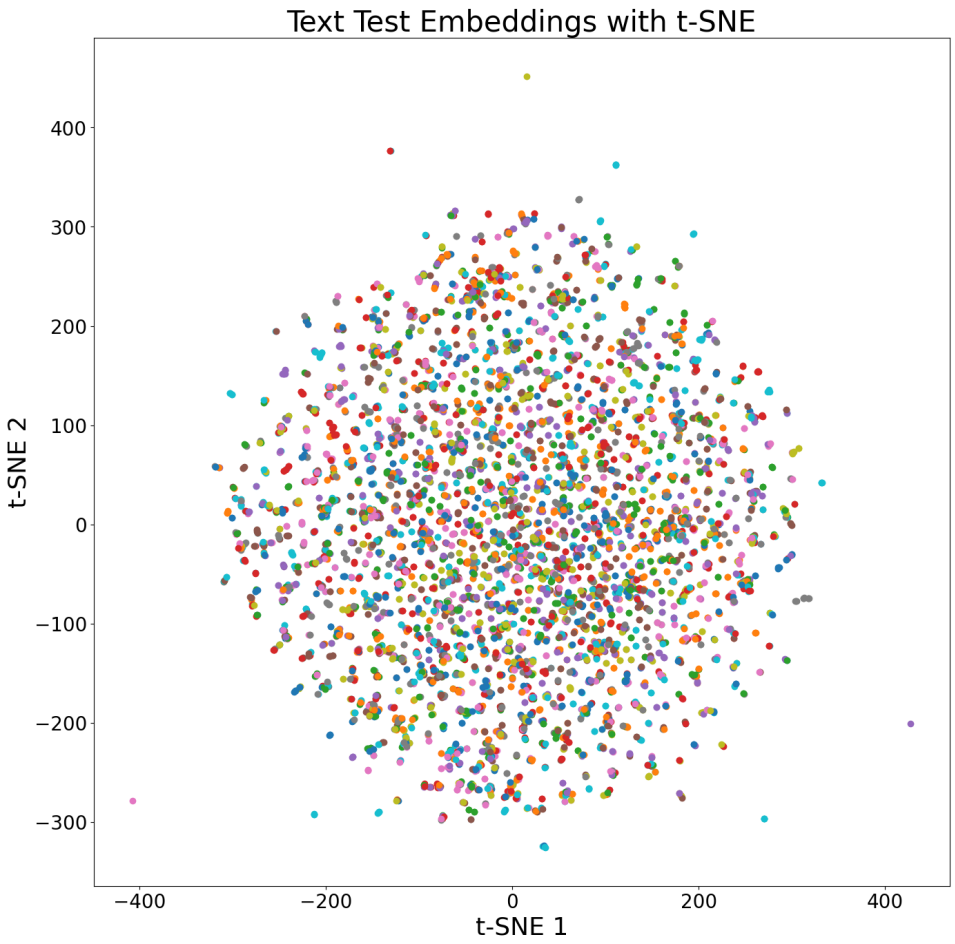


Figure 5. t-SNE for text embedding

The main issue with classifying with product titles is the lack of information, they capture a general description of the product but many details are left out. We have visualized two common examples of when text embeddings are not able to accurately classify matches (See Figure 6 in the Appendix). In the first example, the title indicates a traditional type of men's shirt but the product image on the left is long-sleeve and the one on the right is short-sleeve, this distinction is not present in the title and therefore cannot be incorporated into the embeddings.

The same issue persists for the second example. We expect this model to have difficulty classifying with high accuracy with the low level of detail contained in product titles.

Image Model:

For the image models, ResNet50 and MobileNet, their performance exceeds the benchmark set by the baseline model. As we can see in Table 2 below, ResNet50 outperformed MobileNet across all metrics. ResNet50 has a deeper architecture with a larger number of convolutional filters, allowing it to capture more complex patterns in the data (Wu et al., 2019). ResNet50 contains more parameters so the network is able to remember the information at a higher resolution compared to a compressed model like MobileNet and can attend to products with minor differences from others.

Model	F1 Score on Validation	F1 Score on Test	Accuracy on Test
Random Guessing	0.0024	0.0023	0.0023
ResNet50	0.4635	0.4545	0.4767
MobileNet	0.3890	0.3770	0.3989

Table 2. Metrics for Image Model

Similar to the text embeddings, there is a lot of ambiguity in the distinction of products based on image embeddings, as seen in Figure 7 below. There are some patterns in the instances where ResNet50 misclassified products, specifically (See Figure 7 in Appendix): products with similar designs but in different colors, products with minor textual differences, and images with similar rich color schemes, etc. In the t-SNE graph, these examples would overlap with the product they are mistaken for because they share so many similarities but are not exact enough to be the same product.

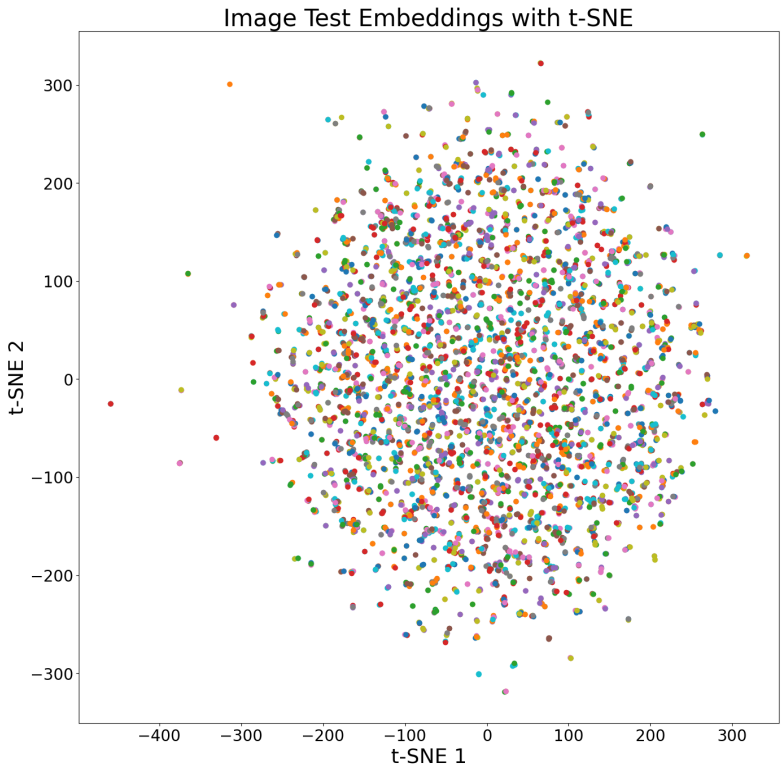


Figure 6. t-SNE for image test data embedding

Image + Text Model:

Model	F1 Score on Validation	F1 Score on Test	Accuracy on Test
DistlBERT + MobileNet	0.5153	0.5025	0.5

Table 3. Metrics for Multimodal Model

Implementing ResNet50 exhausted computational resources, so joint embeddings were generated by previous iterations of MobileNet and DistilBert to make predictions. The multimodal model outperformed the text and image models on the validation and test data by a wide margin, indicating that the emulsion of text and image data can identify more accurate

product matches. This model has slightly higher recall compared to precision which is important for generating similar recommendations because you want to predict the most relevant searches and false positives (bad recommendations) should be avoided (See Figure 3 in Appendix). Product titles lacked detail to make exact matches, however, with information from product images, any gaps in the description are filled with information from image embeddings.

Fine-tuning is not always necessary with pre-trained models initially performing well on desired tasks. In our case, we chose not to fine-tune our training data because the original models were exposed to diverse sets of images and languages which would be useful for generating future matches from products we haven't seen. This prevented the model from overfitting to the training data we worked with, so products outside of the current catalog can be encoded in future iterations of the model.

Given the data is very unbalanced, the final models did not exhibit high accuracy. Multimodal models built with the goal of optimization are able to generate accurate matches 79.3% [2]of the time. But our question of interest was what mode of data provided the most information useful for product matching. Based on the results of our analysis, joint information from product descriptions and images has a better likelihood of capturing similarities in products.

[2] https://www.kaggle.com/competitions/shopee-product-matching/discussion/238136

Conclusion

In conclusion, our project aimed at improving the product matching accuracy for Shopee by developing unimodal (image and text only) and multimodal (image-text) models. Our project faced significant challenges due to the inherent complexity of the problem we set out to solve. The results show that our multimodal model outperforms the unimodal models (SBERT, DistilBERT, ResNet50, MobileNet) in terms of F1 score and accuracy. Our final image-text model leverages image and text data, and classified matches amongst the test set with 53.4% accuracy. The use of both modes of data allows for information that is lacking in product titles to be filled in by image data and vice versa. Our model shows the potential of simplifying the product matching process and attracting more users to the platform, which can ultimately increase Shopee's market share and make it more competitive in the market.

For further implementations, more effort could be applied to incorporating more product features like its p-hash to identify duplicate products based on similarity. In addition, we could upgrade our pipeline to first distinguish them into subgroups or general clusters, such as clothes, pants, shoes, books, electronics, and so on, the same as what we can see in other popular platforms. It would dramatically increase the quality of training data and feature extractions by reducing the noise. Then it would theoretically lead to more accurate prediction results with less usage of computational resources to train.

Roles

Name	Role Description	Peer Review
Suzy & Dingkun	EDA	Isha & Alisa
Isha	Cleaning Text Data	Alisa
Dingkun	Preprocessing Image Data	Suzy
Alisa, Suzy & Isha	Model Building	Dingkun & Isha
Dingkun & Alisa	Model Tuning	Suzy & Isha

References

Baltrusaitis, T., Ahuja, C., Morency, L. P., & Neverova, N. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443. https://doi.org/10.1109/tpami.2018.2798607

Bast, S., Brosch, C., & Krieger, R. (2022). A hybrid approach for product classification based on image and text matching. *Proceedings of the 11th International Conference on Data Science, Technology and Applications*. https://doi.org/10.5220/0011260200003269

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. *O'Reilly Media, Inc.*

Chandel, P. (2021, April 23). *Ensemble of multiple models(lb0.733)*. Kaggle. Retrieved April 19, 2023, from https://www.kaggle.com/code/prashantchandel1097/ensemble-of-multiple-models-lb0-733

Dane, S., Howard, A., Liew, C., & Wong, M. (2021). Shopee - Price match guarantee [Data set]. *Kaggle.* Retrieved from https://kaggle.com/competitions/shopee-product-matching

Estrada-Valenciano, R.; Muñiz-Sánchez, V.; De-la-Torre-Gutiérrez, H. An entity-matching system based on multimodal data for two major E-Commerce stores in mexico. Mathematics 2022, 10, 2564. https://doi.org/10.3390/math10152564

Fang, Y., Wang, J., Jia, L., & Kim, F. W. (2018). *Shopee price Match Guarantee algorithm based on multimodal learning ...* Shopee Price Match Guarantee Algorithm based on multimodal learning. Retrieved April 19, 2023, from https://ieeexplore.ieee.org/abstract/document/9574565/

Gaubys, J. (n.d.). Global ecommerce sales growth (2021–2026) [Apr 2023 update]. Oberlo. Retrieved April 16, 2023, from https://www.oberlo.com/statistics/global-ecommerce-sales-growth

González‐Ibáñez, R., Esparza‐Villamán, A., Vargas‐Godoy, J. C., & Shah, C. (2019). A comparison of Unimodal and multimodal models for implicit detection of relevance in interactive Ir. *Journal of the Association for Information Science and Technology*, *70*(11), 1223–1235. https://doi.org/10.1002/asi.24202

Gupte, K., Pang, L., Vuyyuri, H., & Pasumarty, S. (2021). Multimodal product matching and category mapping: Text+image based Deep Neural Network. *2021 IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/bigdata52589.2021.9671384

Jing Huang, Suya You, & Jiaping Zhao. (2011). Multimodal Image Matching Using Self Similarity. *2011 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. https://doi.org/10.1109/aipr.2011.6176359

Jia, S., Li, C., Yang, L., & Lu, X. (2019). Deep Learning for Hyperspectral Image Classification: An Overview. IEEE Access, 7, 79480-79509. https://doi.org/10.1109/ACCESS.2019.2923803

Kaspien, E. R. C. I. at. (2022, May 9). *Amazon Fair Pricing Policy: Which marketplaces does Amazon price match?* We Help Brands Sell on Amazon, Walmart, eBay & Beyond - Kaspien, Inc. Retrieved April 17, 2023, from https://www.kaspien.com/blog/amazon-fair-pricing-policy-which-marketplaces-does-amazon-price-match/

Łukasik, S., Michałowski, A., Kowalski, P. A., & Gandomi, A. H. (2021). Text-based product matching with incomplete and inconsistent items descriptions. *Computational Science – ICCS 2021*, 92–103. https://doi.org/10.1007/978-3-030-77964-1_8

Nils Reimers, & Iryna Gurevych. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.

Sharma, N., Jain, V., & Mishra, A. (2018). An analysis of convolutional neural networks for image classification. Procedia Computer Science, 132, 377–384. doi:10.1016/j.procs.2018.05.198

Shazia, A., Xuan, T. Z., Chuah, J. H., Usman, J., Qian, P., & Lai, K. W. (2021). A comparative study of multiple neural network for detection of COVID-19 on chest X-ray. EURASIP Journal on Advances in Signal Processing, 2021(1). https://doi.org/10.1186/s13634-021-00755-1

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020, March 1). *Distilbert, a distilled version Bert: Smaller, faster, cheaper and lighter*. arXiv.org. Retrieved April 16, 2023, from https://arxiv.org/abs/1910.01108

Wang, W., Li, Y., Zou, T., Wang, X., You, J., & Luo, Y. (2020). A novel image classification approach via dense-mobilenet models. *Mobile Information Systems*, *2020*, 1–8. https://doi.org/10.1155/2020/7602384

Wang, X., Pang, K., Zhou, X., Zhou, Y., Li, L., & Xue, J. (2015). A visual model-based perceptual image hash for content authentication. IEEE Transactions on Information Forensics and Security, 10(7), 1336–1349. doi:10.1109/TIFS.2015.2407698

Wilame, Williame. (2020, June 15). Why You Should Not Trust Only in Accuracy to Measure Machine Learning Performance. Medium. https://medium.com/@limavallantin/why-you-should-not-trust-only-in-accuracy-to-measure-machine-learning-performance-a72cf00b4516

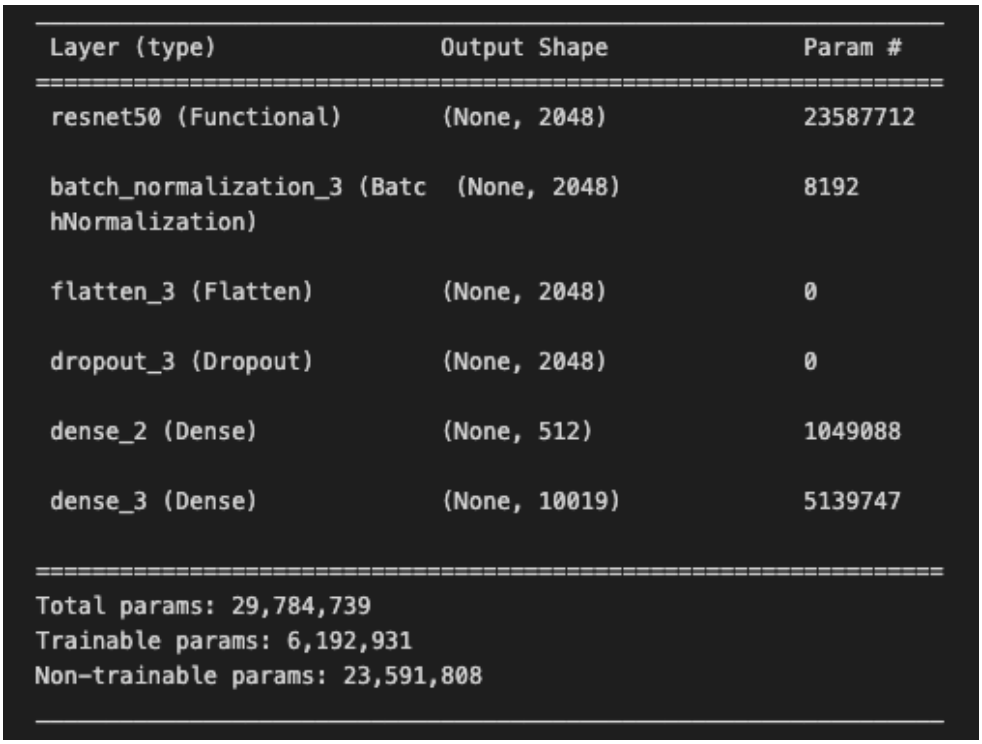
Wu, Z., Shen, C., & van den Hengel, A. (2019). Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. Pattern Recognition, 90, 119–133. https://doi.org/10.1016/j.patcog.2019.01.006

Xie, Z., Liu, L., Wu, Y., Zhong, L., & Li, L. (2021, October 22). Learning text-image joint embedding for efficient cross-modal retrieval with Deep Feature Engineering. arXiv.org. Retrieved April 16, 2023, from https://arxiv.org/abs/2110.11592

zzy990106. (2021, March 15). *B0+Bert CV0.9*. Kaggle. Retrieved April 19, 2023, from https://www.kaggle.com/code/zzy990106/b0-bert-cv0-9

Appendix

Figure 1: ResNet50 model architecture



Layer (type)	Output Shape	Param #
resnet50 (Functional)	(None, 2048)	23587712
batch_normalization_3 (BatchNormalization)	(None, 2048)	8192
flatten_3 (Flatten)	(None, 2048)	0
dropout_3 (Dropout)	(None, 2048)	0
dense_2 (Dense)	(None, 512)	1049088
dense_3 (Dense)	(None, 10019)	5139747

Total params: 29,784,739
Trainable params: 6,192,931
Non-trainable params: 23,591,808

Figure 2: Number of Tokens in Training Data Product Titles

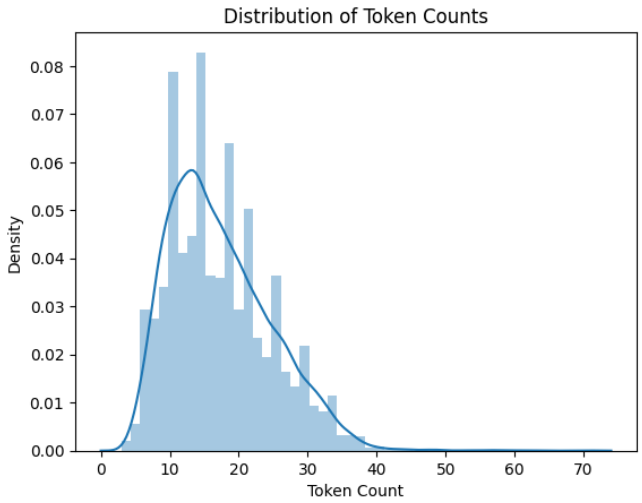


Figure 3: Classification Report from DistlBERT + MobileNet Multimodal Model

	precision	recall	f1-score	support
258047	0.00	0.00	0.00	1
645628	0.00	0.00	0.00	5
887886	1.00	1.00	1.00	1
2942125	0.00	0.00	0.00	0
3108272	0.00	0.00	0.00	1
5029586	0.00	0.00	0.00	0
5488150	1.00	1.00	1.00	1
5949579	0.00	0.00	0.00	1
6381662	0.00	0.00	0.00	1
8297881	1.00	1.00	1.00	1
8660034	0.00	0.00	0.00	1
9079959	1.00	1.00	1.00	1
11497208	1.00	1.00	1.00	1
12491276	1.00	1.00	1.00	1
12910319	1.00	1.00	1.00	1
13645363	0.00	0.00	0.00	0
15630034	0.50	1.00	0.67	1
16448490	1.00	1.00	1.00	4
16856752	1.00	0.50	0.67	2
16933527	1.00	1.00	1.00	3
16993220	0.00	0.00	0.00	2
17378411	1.00	1.00	1.00	1
18118282	0.00	0.00	0.00	0
...				
accuracy			0.53	6850
macro avg	0.37	0.40	0.37	6850
weighted avg	0.51	0.53	0.50	6850

Figure 4: KNN Validation for Text Models

DistilBERT *SBERT*

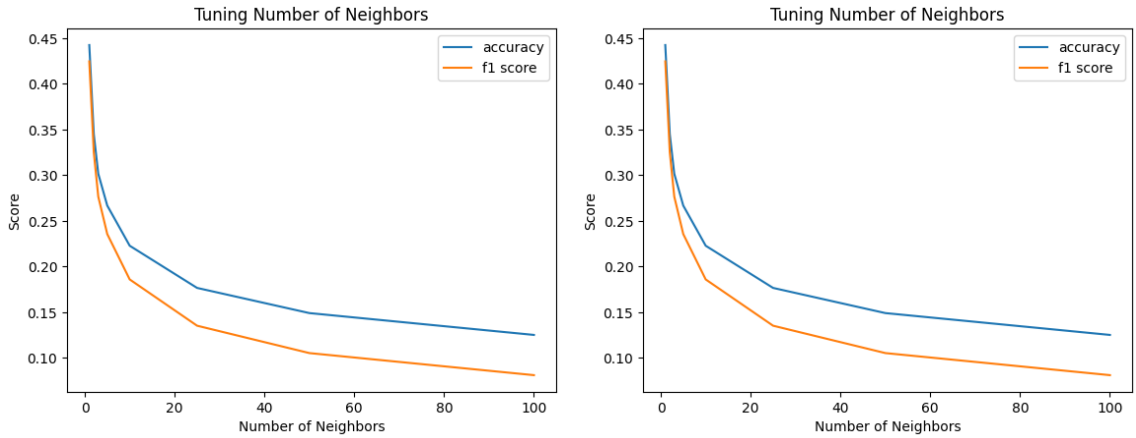


Figure 5: KNN Validation on Image Models

MobileNet *ResNet50*

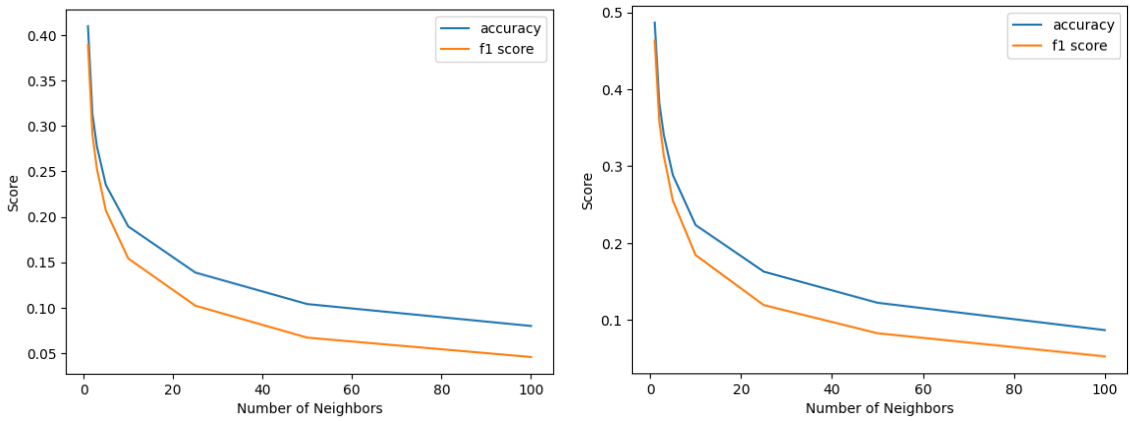


Figure 6: Common misclassified products by Text Models

Baju Koko Pria Gus Azmi Syubbanul Muslimin Kombinasi Hadroh Azzahir Hilw HO187 KEMEJA KOKO PRIA BAJU

Koko syubbanul muslimin koko azzahir koko baju

Example 2.1: The products are very similar (Mens' formal shirts), however they have slight differences like sleeve length and pattern.

Example 2.2: Both products are adapters for Android chargers, however they are slightly different in their input cable. This could be overlooked since it is not specified in the title itself.

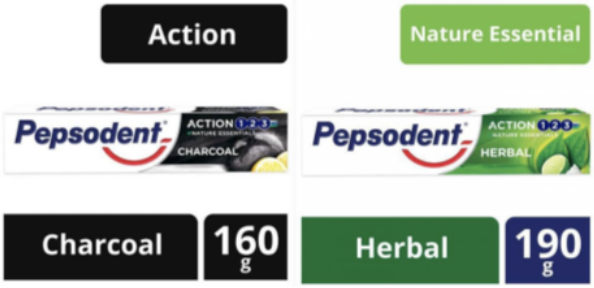
OTG Mini Micro USB Adapter Handphone Hp Smartphone On The Go Adaptor

OTG CONNECTOR MOBILE MURAH MICRO ANDROID

We found that DistillBERT has poor performance as indicated in the table:

1. Products with nuance in text
2. Title in foreign languages

Figure 7: Common misclassified products by Image Models



Example 1: Images with similar design but differences in color and text were misclassified

Example 2: Similar products in the background, wrong image was attended to and misclassified.

Example 3: Almost identical products but there are differences in text.

We found that MobileNet has poor performance as indicated in the pictures:

1. Products with similar design but nuance in texts
2. Products with similar colors

Multimodal model:

Table 1: Performance of text-image model on previous examples:

	Classification
Example 1	Wrong
Example 2	Wrong (with different predicted class)
Example 3	Correct
Example 4	Wrong
Example 5	Correct